(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 April 2001 (12.04.2001)

PCT

(10) International Publication Number
WO 01/25417 A2

(51) International Patent Classification[7]: C12N 15/10, C12Q 1/68

(21) International Application Number: PCT/GB00/03765

(22) International Filing Date: 2 October 2000 (02.10.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

9923327.2	1 October 1999 (01.10.1999)	GB
0011068.4	8 May 2000 (08.05.2000)	GB
0013106.0	30 May 2000 (30.05.2000)	GB

(71) Applicant *(for all designated States except US)*: **GENDAQ LIMITED** [GB/GB]; 1-3 Burtonhole Lane, Mill Hill, London NW7 1AD (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **CHOO, Yen** [GR/GB]; MRC Laboratory of Molecular Biology, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH (GB). **KLUG, Aaron** [GB/GB]; 70 Cavendish Avenue, Cambridge CB1 7UT (GB).

(74) Agents: MASCHIO, Antonio et al.; D Young & Co, 21 New Fetter Lane, London EC4A 1DA (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/25417 A2

(54) Title: DNA LIBRARY

(57) Abstract: A library is provided of DNA sequences consisting of 4^N sequences, where N is greater than or equal to three, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of a DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate. Preferably, each sequence occupies a discrete position, and preferably, the library is arranged in two or more sub-libraries, preferably 4N sub-libraries, wherein for any one sub-library one base in the DNA sequence of length N is defined and the other N-1 bases are randomised. The library may be used in screening methods to identify and characterise zinc fingers having specificity for particular nucleotide sequences.

DNA LIBRARY

Field of the Invention

The present invention relates to a library of DNA sequences immobilised onto a solid support and its use in methods of selecting and designing polypeptides comprising nucleic acid binding motifs, in particular zinc finger polypeptides.

Background of the Invention

Selective gene expression is mediated via the interaction of protein transcription factors with specific nucleotide sequences within the regulatory region of the gene. The most widely used domain within protein transcription factors appears to be the zinc finger (Zf) motif. This is an independently folded zinc-containing mini-domain which is used in a modular repeating fashion to achieve sequence-specific recognition of DNA. The first zinc finger motif was identified in the *Xenopus* transcription factor TFIIIA. The structure of Zf proteins has been determined by NMR studies (Lee *et al.*, 1989 Science 245, 635-637) and crystallography (Pavletich & Pabo, 1991, Science 252, 809-812).

The manner in which DNA-binding protein domains are able to discriminate between different DNA sequences is an important question in understanding crucial processes such as the control of gene expression in differentiation and development. The zinc finger motif has been studied extensively, with a view to providing some insight into this problem, owing to its remarkable prevalence in the eukaryotic genome, and its important role in proteins which control gene expression in *Drosophila*, mice and humans (Kinzler *et al.*, 1988 Nature (London) 332, 371).

Most sequence-specific DNA-binding proteins bind to the DNA double helix by inserting an α-helix into the major groove. Sequence specificity results from the geometrical and chemical complementarity between the amino acid side chains of the and the accessible groups exposed on the edges of base-pairs. In addition to this direct reading of the DNA

sequence, interactions with the DNA backbone stabilise the complex and are sensitive to the conformation of the nucleic acid, which in turn depends on the base. *A priori*, a simple set of rules might suffice to explain the specific association of protein and DNA in all complexes, based on the possibility that certain amino acid side chains have preferences for particular base-pairs. However, crystal structures of protein-DNA complexes have shown that proteins can be idiosyncratic in their mode of DNA recognition, at least partly because they may use alternative geometries to present their sensory α-helices to DNA, allowing a variety of different base contacts to be made by a single amino acid and vice versa (Matthews 1988 Nature (London) 335, 294-295).

Mutagenesis of Zf proteins has confirmed modularity of the domains. Site directed mutagenesis has been used to change key Zf residues, identified through sequence homology alignment, and from the structural data, resulting in altered specificity of Zf domain (Nardelli *et al.*, 1992 NAR 26, 4137-4144). The authors suggested that although design of novel binding specificities would be desirable, design would need to take into account sequence and structural data. They state "there is no prospect of achieving a zinc finger recognition code".

Despite this, many groups have been trying to work towards such a code, although only limited rules have so far been proposed. For example, Desjarlais *et al.*, (1992b PNAS 89, 7345-7349) used systematic mutation of two of the three contact residues (based on consensus sequences) in finger two of the polypeptide Sp1 to suggest that a limited degenerate code might exist. Subsequently the authors used this to design three Zf proteins with different binding specificities and affinities (Desjarlais & Berg, 1993 PNAS 90, 2250-2260). They state that the design of Zf proteins with predictable specificities and affinities "may not always be straightforward".

The crystal structures of zinc finger-DNA complexes show a semiconserved pattern of interactions in which 3 amino acids from the α-helix contact 3 adjacent bases (a triplet) in DNA (Pavletich & Pabo 1991 Science 252, 809-817; Fairall *et al.*, 1993 Nature (London) 366, 483-487; and Pavletich & Pabo 1993 Science 261, 1701-1707). Thus the mode of DNA recognition is principally a one-to-one interaction between amino acids and bases. Because

zinc fingers function as independent modules, it should be possible for fingers with different triplet specificities to be combined to give specific recognition of longer DNA sequences. Each finger is folded so that three amino acids are presented for binding to the DNA target sequence, although binding may be directly through only two of these positions. In the case of Zif268 for example, the protein is made up of three fingers which contact a 9 base pair contiguous sequence of target DNA. A linker sequence is found between fingers which appears to make no direct contact with the nucleic acid.

Protein engineering experiments have shown that it is possible to alter rationally the DNA-binding characteristics of individual zinc fingers when one or more of the α-helical positions is varied in a number of proteins (Nardelli *et al.*, 1991, Nature (London) 349, 175-178; Nardelli *et al.*, 1992, Nucleic Acids Res. 20, 4137-4144; and Desjarlais & Berg 1992a, Proteins 13, 272). It has already been possible to propose some principles relating amino acids on the α-helix to corresponding bases in the bound DNA sequence (Desjarlais & Berg 1992b, Proc. Natl. Acad. Sci. USA 89, 7345-7349). However in this approach the altered positions on the α-helix are prejudged, making it possible to overlook the role of positions which are not currently considered important; and secondly, owing to the importance of context, concomitant alterations are sometimes required to affect specificity (Desjarlais & Berg 1992b), so that a significant correlation between an amino acid and base may be misconstrued.

To investigate binding of mutant Zf proteins, Thiesen and Bach (1991 FEBS 283, 23-26) mutated Zf fingers and studied their binding to randomised oligonucleotides, using electrophoretic mobility shift assays. Subsequent use of phage display technology has permitted the expression of random libraries of Zf mutant proteins on the surface of bacteriophage. The three Zf domains of Zif268, with 4 positions within finger one randomised, have been displayed on the surface of filamentous phage by Rebar and Pabo (1994 Science 263, 671-673). The library was then subjected to rounds of affinity selection by binding to target DNA oligonucleotide sequences to obtain Zf proteins with new binding specificities. Randomised mutagenesis (at the same postions as those selected by Rebar &

Pabo) of finger 1 of Zif 268 with phage display has also been used by Jamieson *et al.*, (1994 Biochemistry 33, 5689-5695) to create novel binding specificity and affinity.

More recently Wu *et al.* (1995 Proc. Natl. Acad. Sci. USA 92, 344-348) have made three libraries, each of a different finger from Zif268, and each having six or seven α-helical positions randomised. Six triplets were used in selections but did not return fingers with any sequence biases; and when the three triplets of the Zif268 binding site were individually used as controls, the vast majority of selected fingers did not resemble the sequences of the wild-type Zif268 fingers and, though capable of tight binding to their target sites *in vitro*, were usually not able to discriminate strongly against different triplets. The authors interpret the results as evidence against the existence of a code.

In summary, it is known that Zf protein motifs are widespread in DNA binding proteins and that binding is via three key amino acids, each one contacting a single base pair in the target DNA sequence. Motifs are modular and may be linked together to form a set of fingers which recognise a contiguous DNA sequence (e.g. a three fingered protein will recognise a 9mer etc). The key residues involved in DNA binding have been identified through sequence data and from structural information. Directed and random mutagenesis has confirmed the role of these amino acids in determining specificity and affinity. Phage display has been used to screen for new binding specificities of random mutants of fingers. A recognition code, to aid design of new finger specificities, has been worked towards although it has been suggested that specificity may be difficult to predict.

Given the lack of predictability in the outcome of rational zinc finger engineering, there is a need for a reliable method for checking the results of efforts to custom design zinc fingers with desired sequence specificity, whether such zinc fingers are obtained by design ("rational design") or by selection from random mutants ("empirical selection"). Not only should the target sequence be included in the test assay but also related sequences because (i) selection is by affinity and not necessarily by specificity and (ii) as discussed, rational design is unreliable owing to degenerate recognition codes, incomplete code and/or unpredictable synergistic contacts.

Ideally, the assay should include all possible DNA sequences, of given length, to establish the preferred specificity of the protein motif to rank other acceptable DNA sequences in terms of affinity. Therefore, wherever possible, an idea of the absolute affinity should emerge in parallel, i.e. the assay should not be simply comparative. This is possible by, for example, determining the apparent Kd of a protein for a series of related binding sites.

However, as the number of test binding sites in the assay increases, it becomes unfeasible to achieve this using prior art techniques. One possible method is to use the SELEX technique (Thiesen and Bach, 1991, FEBS 283, 23-26). However this technique is (i) iterative and hence laborious, (ii) comparative not quantitative, no K_ds emerge, (iii) requires empirical determination of starting parameters and (iv) if selection rounds are carried too far then all comparative information is lost too, as only the best site survives the selection. In addition, as selection is exponential (by PCR) very small differences in DNA-binding preferences can result in apparently huge selection pressures.

Summary of the Invention

We have found that using DNA chip technology to immobilise all the necessary DNA sequences onto a solid phase format allows improved selection for zinc fingers with particular sequence specificity. Since at each stage of the selection procedure, all possible binding sites are present, specificity can be easily confirmed.

Accordingly, the present invention provides a library of DNA sequences consisting of 4^N sequences, where N is greater than or equal to three, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of a DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate.

The present invention also provides a method for designing a zinc finger polypeptide having specificity for a particular DNA sequence comprising a contiguous sequence of N nucleotides, where N is greater than or equal to three, which method comprises:

(i) providing a zinc finger polypeptide, preferably by designing using a rational design method or by selection from a library;

(ii) producing the polypeptide;

(iii) determining the sequence specificity for the polypeptide by contacting a library of DNA sequences with the polypeptide and identifying the sequence or sequences with which the polypeptide binds to with greatest affinity;

(iv) if the sequence or sequences identified in step (iii) are not the desired sequences, making modifications to the amino acid sequence of the polypeptide, preferably based on rational design or by selection from a library, and repeating steps (ii) and (iii).

wherein the library of DNA sequences consist of 4^N sequences, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of the DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate.

The present invention also provides a method for isolating a zinc finger polypeptide having specificity for a particular DNA sequence comprising a contiguous sequence of N nucleotides, where N is greater than or equal to three, which method comprises:

(i) contacting a library of carrier organisms which express on their surface a zinc finger polypeptide comprising variations in the amino acid sequence of the zinc finger DNA binding domain, with a library of DNA sequences; and

(ii) selecting those carrier organisms which express a zinc finger polypeptide that binds to the particular DNA sequence; and

(iii) optionally repeating selection steps (i) and (ii) with those carrier organisms selected in step (ii),

wherein the library of DNA sequences consist of 4^N sequences, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of the DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate.

In another aspect the present invention provides a method for determining the preferred base recognition specificity of a zinc finger polypeptide, which method comprises contacting a library of DNA sequences with the polypeptide, measuring the affinity with which the polypeptide binds to each of the sequences, and optionally ranking the sequences in order of the affinity with which the polypeptide binds,

wherein the library of DNA sequences consist of 4^N sequences, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of the DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate.

In a preferred embodiment of the invention, each of the DNA sequences within the library occupies a discrete position on the solid substrate.

The present invention also provides the use of a library of the invention in a method for designing a zinc finger polypeptide having specificity for a particular DNA sequence.

The present invention further provides the use of a library of the invention in a method for isolating a zinc finger polypeptide having specificity for a particular DNA sequence.

The present invention additionally provides the use of a library of the invention in a method for determining the preferred base recognition specificity of a zinc finger polypeptide.

The DNA library may be arranged into two or more sub-libraries. Each sub-library may occupy a discrete position on the solid substrate. Preferably, each sub-library comprises a subset of the 4^N sequences. In a preferred embodiment of the invention, the library is arranged in 4N sub-libraries, wherein for any one sub-library one base in the DNA sequence of length N is defined and the other N-1 bases are randomised. According to a further aspect of the invention, we provide such a sub-library.

Brief Description of the Drawings

Fig 1. Overview of the protein engineering strategy.

Step 1. Two pre-made zinc finger phage-display libraries, Lib12 and Lib23, contain randomised DNA-binding amino acid positions in fingers 1 and 2 (black) or fingers 2 and 3 (grey) respectively. Selections of 'one-and-a-half' fingers from each master library are carried out in parallel using DNA sequences in which 5 nucleotides have been fixed to a sequence of interest.

Step 2. Zinc finger genes are amplified from the recovered phage using PCR and sets of 'one-and-a-half' fingers are paired to yield recombinant three-finger DNA-binding domains.

Step 3. The recombinant DNA-binding domains are cloned back into phage and subjected to further rounds of selection, or immediately validated for binding to a composite 10 bp DNA of pre-defined sequence.

Fig 2. Composition of the 'bipartite' library.

(a) DNA recognition by the two zinc finger master libraries, Lib12 and Lib23. The libraries are based on the three-finger DNA-binding domain of Zif268 and the putative binding scheme is based on the crystal structure of the wild-type domain in complex with DNA (Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**, 809-817 (1991); Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. Zif268 protein-DNA complex refined at 1.6A: a model system for understanding zinc finger interactions. *Structure* **4**, 1171-1180 (1996)). The DNA-binding positions of each zinc finger are numbered and randomised residues in the two libraries are circled. Broken arrows denote possible DNA contacts from Lib12 to bases H'IJKLM and from Lib23 to bases MNOPQ. Solid arrows show DNA contacts from those regions of the two libraries that carry the wild-type Zif268 amino acid sequence, as observed in the crystal structure. The wild-type portion of each library target site (white boxes) determines the register of the zinc finger-DNA interactions, such that the selected portions of the two libraries can be recombined to recognise the composite site H'IJKLMNOPQ.

(b) Amino acid composition of the randomised DNA-binding positions on the α-helix of each zinc finger. A subset of the 20 amino acids was included in each DNA-binding position. Note that positions 4 and 5 of F2 (LS) are specified by the codons CTG AGC, which contain the recognition site of the restriction enzyme *Dde*I (underlined), used as a breakpoint to recombine the products of the two libraries.

Table 1. Selection of DNA-binding domains to recognise the HIV-1 promoter.

(a) Nucleotide sequences from HIV-1 of the form 3'-HIJKLMNOPQ-5' as recognised by phage clones A-G. Bases which are predicted to be bound by amino acid residues from Lib12 and Lib23, according to the model described in Fig. 2, are shown in bold black and grey, respectively. The position of base Q in each site is numbered relative to the transcription start site (+1) in the HIV promoter. Note that the binding site for Clone A contains 5 bases from the binding site of Zif268 (underlined); and that this clone was thus derived directly from Lib23, without the need for recombination.

(b) Amino acid sequences of the helical regions from recombinant zinc finger DNA-binding domains that recognise HIV-1 sequences. The origin of the amino acids is indicated by shading Lib12 and Lib23 residues in bold black and grey, respectively. Clone A, which was derived solely from Lib23, contains wild-type Zif268 residues (underlined).

(c) Apparent K_d for the interaction of the customised DNA-binding domains for their cognate sequences as measured by phage ELISA.

Figure 3 shows a matrix specificity assay for seven zinc finger DNA-binding domains designed to bind sequences in the HIV-1 promoter. The seven constructs and their respective binding sites are labelled A-G. Binding of zinc fingers to 0.4 pmol DNA per 50μl well is plotted vertically from phage ELISA absorbance readings (A_{450}-A_{650}). Each clone is tested using all seven DNA sequences but strong binding is only observed to those sequences against which they have been designed.

Detailed description of the invention

Although we have described the libraries and methods of our invention with reference to the selection, design, etc of a zinc finger polypeptide, it will be understood that our invention may be applied to other DNA or nucleic acid binding molecules, such as nucleic acid binding proteins or polypeptides (e.g., helix-turn-helix proteins), other nucleic acids such as DNA, RNA, or PNA (protein-nucleic acid), small molecules such as drug, an intercalating molecule, a major or minor groove binding molecule (such as distamycin), etc.

Thus, in a broad sense, our invention encompasses libraries and methods for designing, isolating, and determining the preferred base recognition specificity of any nucleic acid binding molecule.

A. DNA library

A DNA library of the invention is used to test the selectivity of a zinc finger for a nucleotide sequences of length N. Consequently, since there are four different nucleotides that occur naturally in genomic DNA, the total number of sequences required to represent all possible base permutations for a sequence of length N is 4^N. However, uracil, which occurs in RNA, or other natural or non natural bases, may also be included, either in substitution for thymidine, or in addition. Thus, the DNA library of the invention may have 5^N sequences.

N is an integer having a value of at least three. That it to say that the smallest library envisaged for testing binding to a nucleotide sequence where only one DNA triplet is varied, consists of 64 different sequences. However, N may be any integer greater than or equal to 3 such as 4, 5, 6, 7, 8 or 9. Typically, N only needs to be three times the number of zinc fingers being tested, optionally including a few additional residues outside of the binding site that may influence specificity. Thus, by way of example, to test the specificity of a protein comprising three zinc fingers, where all three fingers have been engineered, it may be desirable to use a library where N is at least 9. The DNA sequences in the library

are typically immobilised at discrete positions on a solid substrate, such as a DNA chip, such that each different sequence is separated from other sequences on the solid substrate.

The 4^N possible permutations of the DNA sequence of length N sequence are typically (but need not be) arranged in sub-libraries. Preferably, the library is sub-divided into 4N sub-libraries, wherein for any one sub-library one base in the DNA sequence of length N is defined and the other N-1 bases are randomised. Thus in the case of a varied DNA triplet, there will be 12 sub-libraries.

The nucleotide sequence of length N may be generally, but need not be, part of a longer DNA molecule. Thus, the DNA sequences within the library may consist of sequences against which the binding of a binding molecule is tested (i.e., every base position in the DNA sequence is potentially involved binding to the binding molecule). An example is a library of 64 sequences of length 3 representing all possible targets for a zinc finger motif.

Alternatively, and preferably, the DNA sequences comprise other flanking sequences which are not directly relevant to or involved in binding. Examples of such sequences include vector sequences, dimerisation sequences, or nucleic acid sequences which are capable of hybridising to other nucleic acid sequences to form double stranded regions, other binding targets, etc. The sequence and (where applicable, the binding specificity) of such flanking regions may be known or unknown.

For example, the DNA sequences may comprise one or more binding targets for another binding domain, whether this is a zinc finger domain or otherwise. Such libraries are useful in designing, isolating, and determining the binding affinity and preferred base recognition specificity of a hybrid binder such as a zinc finger-homeodomain fusion protein.

The nucleotide sequence of length N typically occupies the same position within the longer molecule in each of the varied sequences even though the sequence of N itself may vary. The other sequences within the DNA molecule are generally the same throughout the

library. Thus the library can be said to consist of a library of 4^N DNA molecules of the formula R^1-[A/C/G/T]$_4^N$-R^2, wherein R^1 and R^2 may be any nucleotide sequence.

Preferably, each sequence is also represented as a dilution/concentration series. Thus the immobilised DNA library may occupy $Z4^N$ discrete positions on the chip where Z is the number of different dilutions in the series and is an integer having a value of at least 2. The range of DNA concentrations for the dilution series is typically in the order of 0.01 to 100 pmol cm^{-2}, preferably from 0.05 to 5 pmol cm^{-2}. The concentrations typically vary 10-fold, i.e. a series may consist of 0.01, 0.1, 1, 10 and 100 pmol cm^{-2}, but may vary, for example, by 2- or 5-fold.

The advantage of including the DNA sequences in a dilution series is that it is then possible to estimate K_ds for protein/DNA complexes using standard techniques such as the Kaleidagraph™ version 2.0 program (Abelback Software).

The DNA molecules in the library are at least partially double-stranded, in particular at least the nucleotide sequence of length N is double-stranded. Single stranded regions may be included, for example to assist in attaching the DNA library to the solid substrate.

Techniques for producing immobilised libraries of DNA molecules have been described in the art. Generally, most prior art methods described how to synthesise single-stranded nucleic acid molecule libraries, using for example masking techniques to build up various permutations of sequences at the various discrete positions on the solid substrate. U.S. Patent No. 5,837,832, the contents of which are incorporated herein by reference, describes an improved method for producing DNA arrays immobilised to silicon substrates based on very large scale integration technology. In particular, U.S. Patent No. 5,837,832 describes a strategy called "tiling" to synthesize specific sets of probes at spatially-defined locations on a substrate which may be used to produced the immobilised DNA libraries of the present invention. U.S. Patent No. 5,837,832 also provides references for earlier techniques that may also be used.

However, an important aspect of the present invention is that it relates to DNA binding proteins, zinc fingers, that bind double-stranded DNA. Thus single-stranded nucleic acid molecule libraries using the prior art techniques referred to above will then need to be converted to double-stranded DNA libraries by synthesising a complementary strand. An example of the conversion of single-stranded nucleic acid molecule libraries to double-stranded DNA libraries is given in Bulyk *et al.*, 1999, Nature Biotechnology 17, 573-577, the contents of which are incorporated herein by reference. The technique described in Bulyk *et al.*, 1999, typically requires the inclusion of a constant sequence in every member of the library (i.e. within R^1 or R^2 in the generic formula given above) to which a nucleotide primer is bound to act as a primer for second strand synthesis using a DNA polymerase and other appropriate reagents. If required, deoxynucleotide triphosphates (dNTPs) having a detectable labeled may be include to allow the efficiency of second strand synthesis to be monitored. Also the detectable label may assist in detecting binding of zinc fingers when the immobilised DNA library is in use.

Alternatively, double-stranded molecules may be synthesised off the solid substrate and each pre-formed sequence applied to a discrete position on the solid substrate. An example of such a method is to synthesis palindromic single-stranded nucleic acids – see U.S. Patent No. 5556752, the contents of which are incorporated herein by reference.

Thus DNA may typically be synthesised *in situ* on the surface of the substrate. However, DNA may also be printed directly onto the substrate using for example robotic devices equipped with either pins or pizo electric devices.

The library sequences are typically immobilised onto or in discrete regions of a solid substrate. The substrate may be porous to allow immobilisation within the substrate or substantially non-porous, in which case the library sequences are typically immobilised on the surface of the substrate. The solid substrate may be made of any material to which polypeptides can bind, either directly or indirectly. Examples of suitable solid substrates include flat glass, silicon wafers, mica, ceramics and organic polymers such as plastics, including polystyrene and polymethacrylate. It may also be possible to use semi-permeable

membranes such as nitrocellulose or nylon membranes, which are widely available. The semi-permeable membranes may be mounted on a more robust solid surface such as glass. The surfaces may optionally be coated with a layer of metal, such as gold, platinum or other transition metal. A particular example of a suitable solid substrate is the commercially available BiaCore™ chip (Pharmacia Biosensors).

Preferably, the solid substrate is generally a material having a rigid or semi-rigid surface. In preferred embodiments, at least one surface of the substrate will be substantially flat, although in some embodiments it may be desirable to physically separate synthesis regions for different polymers with, for example, raised regions or etched trenches. The solid substrate may be a microtitre plate or bead. It is also preferred that the solid substrate is suitable for the high density application of DNA sequences in discrete areas of typically from 50 to 100 μm, giving a density of 10000 to 40000 cm^{-2}.

The solid substrate is conveniently divided up into sections. This may be achieved by techniques such as photoetching, or by the application of hydrophobic inks, for example teflon-based inks (Cel-line. USA). Where the solid substrate is a microtitre plate, the sections may conveniently comprise the wells of the microtitre plate. Each well may comprise a discrete DNA sequence of the library, or, in the case where the library is sub-divided into sub-libraries, each well may comprise one or more sub-libraries.

Discrete positions, in which each different member of the library is located may have any convenient shape, e.g., circular, rectangular, elliptical, wedge-shaped, etc. A discrete position is commonly referred to as a "spot". Each discrete position may comprise, preferably consist of, one DNA sequence of the library. Thus, the discrete position may comprise a single molecule, or a number of DNA molecules of homogenous composition. The latter arrangement is advantageous in that the signal strength is likely to be higher.

In an alternative embodiment, each discrete position comprises a number of DNA molecules of heterogenous composition. In this embodiment, a number of different DNA sequences are immobilised at a discrete spot. Where the library is divided into sub-

libraries, as described above, preferably each discrete spot comprises the sequences within the sub-library. Thus, in a preferred embodiment, where the library is sub-divided into 4N sub-libraries, each of the sub-libraries is immobilised in a discrete position on the solid substrate. This embodiment is referred to as "multiplexing".

Attachment of the library sequences to the substrate may be by covalent or non-covalent means. The library sequences may be attached to the substrate via a layer of molecules to which the library sequences bind. For example, the library sequences may be labelled with biotin and the substrate coated with avidin and/or streptavidin. A convenient feature of using biotinylated library sequences is that the efficiency of coupling to the solid substrate can be determined easily. Since the library sequences may bind only poorly to some solid substrates, it is often necessary to provide a chemical interface between the solid substrate (such as in the case of glass) and the library sequences. Examples of suitable chemical interfaces include hexaethylene glycol. Another example is the use of polylysine coated glass, the polylysine then being chemically modified using standard procedures to introduce an affinity ligand. Other methods for attaching molecules to the surfaces of solid substrate by the use of coupling agents are known in the art, see for example WO98/49557.

Binding of zinc fingers to the immobilised DNA library may be determined by a variety of means such as changes in the optical characteristics of the bound DNA (i.e. by the use of ethidium bromide) or by the use of labelled zinc finger polypeptides, such as epitope tagged zinc finger polypeptides or zinc finger polypeptides labelled with fluorophores such as green fluorescent protein. Other detection techniques that do not require the use of labels include optical techniques such as optoacoustics, reflectometry, ellipsometry and surface plasmon resonance (SPR) – see WO97/49989, incorporated herein by reference.

Binding of epitope tagged zinc finger polypeptides is typically assessed by immunological detection techniques where the primary or secondary antibody comprises a detectable label. A preferred detectable label is one that emits light, such as a fluorophore, for example phycoerythrin.

The complete DNA library is typically read at the same time by charged coupled device (CCD) camera or confocal imaging system. Alternatively, the DNA library may be placed for detection in a suitable apparatus that can move in an x-y direction, such as a plate reader. In this way, the change in characteristics for each discrete position can be measured automatically by computer controlled movement of the array to place each discrete element in turn in line with the detection means.

The detection means are capable of interrogating each position in the library array optically or electrically. Examples of suitable detection means include CCD cameras or confocal imaging systems.

Any of the immobilised DNA sequences of the library may be removed from the solid substrate for further manipulation. Thus, it may be desired to remove a particular DNA sequence which shows binding to a particular zinc finger, for example. Removal from the solid substrate may be achieved by various means, for example, by elution using an appropriate solvent, by chemical or enzymatic cleavage, photochemical lysis (e.g., by application of laser energy), etc. The removed sequence may be amplified by PCR, for example.

B. Zinc fingers

A zinc finger binding motif is the α-helical structural motif found in zinc finger binding proteins, well known to those skilled in the art. The amino acid numbering used throughout is based on the first amino acid in the α-helix of the zinc finger binding motif being position +1. It will be apparent to those skilled in the art that the amino acid residue at position -1 does not, strictly speaking, form part of the α-helix of the zinc binding finger motif. Nevertheless, the residue at -1 is shown to be very important functionally and is therefore considered as part of the binding motif α-helix for the purposes of the present invention.

The zinc finger polypeptide sequences to be tested and/or selected using the methods of the invention are typically obtained by modifying one or more amino acids residues known to be

important in binding specificity. Thus, for example, zinc finger polypeptide sequences may comprise a substitution at one or more of the following positions: –1, +1, +2, +3, +5 +6 and +8.

Zinger finger polypeptides may in one embodiment be tested individually using the library and methods of the invention. For example, it may be desired to determine the preferred base recognition specificity of a zinc finger polypeptide designed using rational design techniques.

The term "rational design" is intended to refer to the design of a zinc finger sequence according to one or more rules (recognition rules). Various rational design techniques and rules are known in the art, for example, as disclosed in WO98/53057. Thus, according to WO98/53057, a zinc finger may be designed to bind to a nucleic acid quadruplet in a target nucleic acid sequence, wherein binding to each base of the quadruplet by an α-helical zinc finger nucleic acid binding motif in the protein is determined as follows: if base 4 in the quadruplet is G, then position +6 in the α-helix is Arg or Lys; if base 4 in the quadruplet is A, then position +6 in the α-helix is Glu, Asn or Val; if base 4 in the quadruplet is T, then position +6 in the α-helix is Ser, Thr, Val or Lys; if base 4 in the quadruplet is C, then position +6 in the α-helix is Ser, Thr, Val, Ala, Glu or Asn; if base 3 in the quadruplet is G, then position +3 in the α-helix is His; if base 3 in the quadruplet is A, then position +3 in the α-helix is Asn; if base 3 in the quadruplet is T, then position +3 in the α-helix is Ala, Ser or Val; provided that if it is Ala, then one of the residues at -1 or +6 is a small residue; if base 3 in the quadruplet is C, then position +3 in the α-helix is Ser, Asp, Glu, Leu, Thr or Val; if base 2 in the quadruplet is G, then position -1 in the α-helix is Arg; if base 2 in the quadruplet is A, then position -1 in the α-helix is Gln; if base 2 in the quadruplet is T, then position -1 in the α-helix is His or Thr; if base 2 in the quadruplet is C, then position -1 in the α-helix is Asp or His; if base 1 in the quadruplet is G, then position +2 is Glu; if base 1 in the quadruplet is A, then position +2 Arg or Gln; if base 1 in the quadruplet is C, then position +2 is Asn, Gln, Arg, His or Lys; if base 1 in the quadruplet is T, then position +2 is Ser or Thr.

These rules permit the design of a zinc finger binding protein specific for any given nucleic acid sequence. It has been found that position +2 in the helix is responsible for determining the binding to base 1 of the quadruplet. In doing so, it cooperates synergistically with position +6, which determines binding at base 4 in the quadruplet, bases 1 and 4 being overlapping in adjacent quadruplets.

Although zinc finger polypeptides are considered to bind to overlapping quadruplet sequences, rational design rules such as the rules set out above allow polypeptides to be designed to bind to target sequences which are not multiples of overlapping quadruplets. For example, a zinc finger polypeptide may be designed to bind to a palindromic target sequence. Such sequences are commonly found as, for example, restriction enzyme target sequences. Furthermore, creation of zinc fingers which bind to fewer than three nucleotides may be achieved by specifying, in the zinc finger, amino acids which are unable to support H-bonding with the nucleic acid in the relevant position. Advantageously, this is achieved by substituting Gly at position -1 (to eliminate a contact with base 2) and/or Ala at positions +3 and/or +6 (to eliminate contacts at the 3rd or 4th base respectively). The contact with the final (3') base in the target sequence may be strengthened, if necessary, by substituting a residue at the relevant position which is capable of making a direct contact with the phosphate backbone of the nucleic acid.

In an alternative embodiment, a library of zinc finger polypeptides having different amino acids at one or more positions involved in binding specificity may be screened ("empirical selection") using the library and methods of the present invention and zinc finger polypeptides selected that bind to a target nucleotide sequence. Such a library of sequences may conveniently be obtained by random mutagenesis at particular positions to produce a phage display library using standard techniques (see WO96/06166 for construction of a randomised Zif268 library).

Where a randomised zinc finger polypeptide library is used, preferably the zinc fingers are randomised at one or more of, or may have a random allocation at some or all, preferably all,

of positions –1, +1, +2, +3, +5 +6, +8 and +9. More preferably, the zinc fingers are randomised at positions –1, +2, +3 and +6, and at least one of +1, +5 and +8.

The sequences may also be randomised at other positions (e.g. at position +9, although it is generally preferred to retain an arginine or a lysine residue at this position). Further, whilst allocation of amino acids at the designated "random" positions may be genuinely random, it is preferred to avoid a hydrophobic residue (Phe, Trp or Tyr) or a cysteine residue at such positions.

Preferably the zinc finger binding motif is present within the context of other amino acids (which may be present in zinc finger proteins), so as to form a zinc finger (which includes an antiparallel β-sheet). Further, the zinc finger is preferably displayed as part of a zinc finger polypeptide, which polypeptide comprises a plurality of zinc fingers joined by an intervening linker peptide. Typically the library of sequences is such that the zinc finger polypeptide will comprise two or more zinc fingers of defined amino acid sequence (generally the wild type sequence) and one zinc finger having a zinc finger binding motif randomised in the manner defined above. It is preferred that the randomised finger of the polypeptide is positioned between the two or more fingers having defined sequence. The defined fingers will establish the "phase" of binding of the polypeptide to DNA, which helps to increase the binding specificity of the randomised finger.

Preferably the sequences encode the randomised binding motif of the middle finger of the Zif268 polypeptide. Conveniently, the sequences also encode those amino acids N-terminal and C-terminal of the middle finger in wild type Zif268, which encode the first and third zinc fingers respectively. In a particular embodiment, the sequence encodes the whole of the Zif268 polypeptide. Those skilled in the art will appreciate that alterations may also be made to the sequence of the linker peptide and/or the β-sheet of the zinc finger polypeptide.

Typically, the randomised sequence encoding zinc finger polypeptides are such that the zinc finger binding domain can be cloned as a fusion with the minor coat protein (pIII) of bacteriophage fd. Conveniently, the encoded polypeptide includes the tripeptide sequence

Met-Ala-Glu as the N terminal of the zinc finger domain, which is known to allow expression and display using the bacteriophage fd system. Desirably the polypeptide library comprises 10^6 or more different sequences (ideally, as many as is practicable).

C. Uses of the DNA library

Design and testing of custom zinc fingers

The immobilised DNA library of the present invention may conveniently be used to verify the results of rationally designing zinc fingers with desired specificity. Typically a zinc finger motif is designed as described above and then produced by recombinant or synthetic means. The zinc finger polypeptide is contacted with the immobilised DNA library and binding detected as described above. The specificity and affinity of the zinc finger for the various sequences in the library can then be determined. If the desired binding is not seen then further modifications may be made to the zinc finger motif and the screening process repeated.

The use of automated peptide synthesisers and detection means together with computer-controlled equipment and software may allow the process to be fully automated such that when given a target sequence and rational design protocol, the process is repeated automatically until the desired result is obtained.

Screening for zinc finger polypeptides having specificity for one or more DNA sequences.

In another approach, a library of zinc finger polypeptides is contacted with the DNA library and the zinc fingers that bind to the target sequence(s) selected. Conveniently, the zinc finger library is in the form of a library of carrier organisms that express on their surface a zinc finger polypeptide. Typical carrier organisms include phage and bacteria.

Alternatively, other means of phenotype-genotype linkage as known in the art may be used. For example, the libraries may be segregated into compartments or microcapsules, as

described in WO99/02671. This document discloses a method for isolating one or more genetic elements encoding a gene product having a desired activity. Genetic elements are first compartmentalised into microcapsules, and then transcribed and/or translated to produce their respective gene products (RNA or protein) within the microcapsules. Alternatively, the genetic elements are contained within a host cell in which transcription and/or translation (expression) of the gene product takes place and the host cells are first compartmentalised into microcapsules. Genetic elements which produce gene product having desired activity are subsequently sorted. Polysome display techniques, such as those disclosed in WO00/27878, may also be applied to the libraries and methods of our invention.

More than one round of selection may take place, for example to confirm that specificity of zinc finger polypeptides selected in any particular round. Desirably at least two, preferably three or more, rounds of screening are performed.

The library of zinc finger polypeptides need not necessarily be completely random but may be partially random, for example at certain positions only. The positions chosen and the range of different amino acids at any given position may be based on rational design principles.

The two methods are not mutually exclusive and may both be used as part of a design and selection strategy. For example, it may be preferred to use the screening method described above as a precursor to the rational design method described above. Thus in a preferred embodiment, that there is a two-step selection procedure: the first step comprising screening each of a plurality of zinc finger binding motifs (typically in the form of a display library), mainly or wholly on the basis of affinity for the target sequence; the second step comprising comparing binding characteristics of those motifs selected by the initial screening step, and selecting those having preferable binding characteristics for a particular DNA triplet.

The non-specific component of all protein-DNA interactions, which includes contacts to the sugar-phosphate backbone as well as ambiguous contacts to base-pairs, is a considerable driving force towards complex formation and can result in the selection of DNA-binding

proteins with reasonable affinity but without specificity for a given DNA sequence. Therefore, in order to minimise these non-specific interactions when designing a polypeptide, selections should preferably be performed with low concentrations of specific binding site in a background of competitor DNA, and binding should desirably take place in solution to avoid local concentration effects and the avidity of multivalent phage for ligands immobilised on solid surfaces.

As a safeguard against spurious selections, the specificity of individual phage should be determined following the final round of selection.

Determining the preferred base recognition specificity of a zinc finger polypeptide

The immobilised DNA library of the present invention may be used in a general sense to determine the preferred base recognition specificity of a zinc finger polypeptide, whether the zinc finger polypeptide be a naturally occurring zinc finger polypeptide, or a fragment thereof comprising a zinc finger motif, a zinc finger polypeptide identified by a screening procedure, such as the screening method of the invention, or a zinc finger obtained by rational design methods.

Typically, the zinc finger polypeptide of interest in contacted with the DNA library as described above and the extent of binding at each position on the immobilised DNA library determined. The results for each different sequence in the library may then be placed in order of the affinity with which the zinc finger polypeptide binds. The resulting ranking will provide a clear indication of the preferred base recognition specificity of the zinc finger polypeptide and may even be used to determine an optimal consensus binding sequence.

Uses of zinc finger motifs designed and/or selected by the methods of the invention

Once suitable zinc finger binding motifs have been identified and obtained, they will advantageously be combined in a single zinc finger polypeptide. Typically this will be

accomplished by use of recombinant DNA technology; conveniently a phage display system may be used.

In a further aspect the invention provides a zinc finger polypeptide designed and/or selected by one or both of the methods defined above. Preferably the zinc finger polypeptide designed by the method comprises a combination of a plurality of zinc fingers (adjacent zinc fingers being joined by an intervening linker peptide), each finger comprising a zinc finger binding motif. Desirably, each zinc finger binding motif in the zinc finger polypeptide has been selected for preferable binding characteristics by the method defined above. The intervening linker peptide may be the same between each adjacent zinc finger or, alternatively, the same zinc finger polypeptide may contain a number of different linker peptides. The intervening linker peptide may be one that is present in naturally-occurring zinc finger polypeptides or may be an artificial sequence. In particular, the sequence of the intervening linker peptide may be varied, for example, to optimise binding of the zinc finger polypeptide to the target sequence.

Where the zinc finger polypeptide comprises a plurality of zinc binding motifs, it is preferred that each motif binds to those DNA triplets which represent contiguous or substantially contiguous DNA in the sequence of interest. Where several candidate binding motifs or candidate combinations of motifs exist, these may be screened against the actual target sequence to determine the optimum composition of the polypeptide. Competitor DNA may be included in the screening assay for comparison, as described above.

It is well within the capability of one of normal skill in the art to design a zinc finger polypeptide capable of binding to any desired target DNA sequence simply by considering the sequence of triplets present in the target DNA and combining in the appropriate order zinc fingers comprising zinc finger binding motifs having the necessary binding characteristics to bind thereto. The greater the length of known sequence of the target DNA, the greater the number of zinc finger binding motifs that can be included in the zinc finger polypeptide. For example, if the known sequence is only 9 bases long then three zinc finger binding motifs can be included in the polypeptide. If the known sequence is 27 bases long then, in theory, up to

nine binding motifs could be included in the polypeptide. The longer the target DNA sequence, the lower the probability of its occurrence in any given portion of DNA.

Moreover, those motifs selected for inclusion in the polypeptide could be artificially modified (e.g. by directed mutagenesis) in order to optimise further their binding characteristics. Alternatively (or additionally) the length and amino acid sequence of the linker peptide joining adjacent zinc binding fingers could be varied, as outlined above. This may have the effect of altering the position of the zinc finger binding motif relative to the DNA sequence of interest, and thereby exert a further influence on binding characteristics.

Generally, it will be preferred to select those motifs having high affinity and high specificity for the target triplet.

Possible uses of suitably designed zinc finger polypeptides are:

a) Therapy (e.g. targeting to double stranded DNA)

b) Diagnosis (e.g. detecting mutations in gene sequences: the present work has shown that "tailor made" zinc finger polypeptides can distinguish DNA sequences differing by one base pair).

c) DNA purification (the zinc finger polypeptide could be used to purify restriction fragments from solution, or to visualise DNA fragments on a gel - for example, where the polypeptide is linked to an appropriate fusion partner, or is detected by probing with an antibody).

In addition, zinc finger polypeptides could even be targeted to other nucleic acids such as single-stranded or double-stranded RNA (e.g. self-complementary RNA such as is present in many RNA molecules) or to RNA-DNA hybrids, which would present another possible mechanism of affecting cellular events at the molecular level.

Examples

These examples show the use of the DNA libraries of the invention in designing and/or isolating a zinc finger polypeptide having a particular DNA sequence specificity, as well as in the determination of the preferred base recognition specificity of a zinc finger polypeptide.

General Materials and Methods for screening procedure using phage library

1. Prepare DNA chips as in Bulyk *et al.*, 1999, *ibid.*

2. Prepare a fresh phage culture for assay by innoculating 2ml of 2xTY containing 15 µg/ml tetracycline with a single bacterial colony and incubating for 8 - 24 hours at 30°C.

3. Block chip surface for 1 hour at 20°C by adding 150 µl PBS containing 4% (w/v) fat-free freeze-dried milk (Marvel).

4. Centrifuge phage cultures from step 2 on a benchtop microfuge for 10 minutes at top speed to obtain clear phage-containing culture supernatant.

5. Prepare 200 µl phage binding mixture for each assay by mixing 20 µl phage supernatant with 180 µl of PBS containing 2% (w/v) fat-free freeze-dried milk (Marvel), 1% (v/v) Tween and 1 µg competitor nucleic acid, e.g. sonicated salmon sperm DNA or poly dIdC depending on the application.

6. Discard blocking mixture from chip and add phage binding mixture to chip. Incubate for up to 1 hour at 20°C.

7. Remove unbound phage by washing chip 7 times with PBS containing 1% (v/v) Tween followed by 3 washes with PBS.

8. Add PBS containing 2% (w/v) fat-free freeze-dried milk (Marvel) and 0.02% (v/v) biotin-conjugated anti-M13 IgG (Pharmacia Biotech). Incubate for 1 hour at 20°C.

9. Remove unbound antibody by washing chip 3 times with PBS containing 0.05% (v/v) Tween-20 followed by 3 washes with PBS.

10. Add a solution of streptavidin-phycoerythrin to the chip. Allow to bind for 15 minutes at room temp. Remove unbound antibody by washing microtitre plate wells 3 times with PBS containing 0.05% (v/v) Tween-20 followed by 3 washes with PBS.

11. Detection protocols as described in Bulyk *et al.*, 1999, *ibid.*

Example 1 - Use of a DNA chip to study a phage display library of the pZif268 middle finger.

The DNA chip used in this protocol has 64 different features which correspond to the 64 possible middle triplets of the Zif268 binding site. Each DNA binding site is applied to the chip at various densities, covering a roughly 100-fold range, from 0.04 to 4 pmol/cm^2. The DNA sequence synthesised on the chip is: 3'-cctggctaactgaactATATATGCG-NNN-GCGATATAT-5'.

This sequence is attached to the chip at the 3' end of the strand, nucleotides shown in lowercase delineate the primer binding site used in Bulyk *et al.*, 1999, *ibid.*

Screening of entire library on a chip

Every member of the Zif268 phage library (as described in Choo and Klug, 1994, Proc Natl Acad Sci U S A 91, 11163-11167) can be screened against every possible binding site to establish whether the phage display library has any limitations on DNA recognition. This helps ascertain the quality of a library. For instance we now know that the Choo and Klug library has certain sequence-restrictions which arise from the synergy of fingers 2 and 3. The overlap restricts binding to middle triplets with 5' G or T - this is discussed fully in

Isalan *et al.*, 1998, Biochemistry 37: 12026-33 and Isalan *et al.*, 1997, Proc Natl Acad Sci U S A. 94: 5617-21.

Experiment: The library is applied onto a chip with the 64 different triplets. Binding is observed only to those triplets with 5' G or T. Triplets with 5' A or C are not bound: it is concluded that the library is limited.

Following the selection process by screening on a chip.

Experiment 1: During the course of phage selections using the triplet TCC, phage returned from individual rounds of selection are applied on the chip. It is noted that the signal for binding to TCC increases in consecutive rounds of selection, but that there is a higher signal for binding to GAC. It is concluded that (since the phage library is not capable of binding to triplets with 5' T) selection using the oligo with middle triplet TCC (5'-tatata-GCG-TCC-GCG-tatata-3'; putative binding site underlined) has selected fingers that bind quite tightly to a frame-shifted sequence on the complementary strand (3'-atatat-CGC-AGG-CGC-atatat-5'; putative binding site underlined). Note that the frameshift means that finger 1 is forced to recognise the triplet GCT rather than GCG, which is suboptimal. When the triplet GAC is offered to these fingers in the context of the correct binding site for fingers 1 and 3, binding is optimal and a higher signal is obtained. When the amino acid sequences of zinc fingers isolated from separate selections using the triplets TCC and GAC are compared it is seen that the same fingers have been isolated, thus confirming the above hypothesis.

Experiment 2: While carrying out phage selections using the triplet GCG, phage returned from individual rounds of selection are applied on the chip. At each round the signal for GCG is seen to increase relative to the other triplets, demonstrating enrichment. By round 3 it is seen that there is appreciable binding to GCG and very little binding to all other triplets, except for binding to GTG which is also seen. It is concluded that 3 rounds of selection are sufficient to eliminate binders of the other 62 triplets. It is also concluded that the selection has either (i) produced fingers which cannot discriminate between GCG and GTG, or (ii) produced a mixed population of fingers some of which bind GCG and others

GTG. To solve these problems the selection is repeated, including a specific competitor to eliminate GTG binders.

Studying sequence specificity and affinity of (individual) clones on a chip.

Experiment 1: After the GCG selection is repeated, including a specific competitor to eliminate GTG binders, two different ZnF clones [α-helix seq (A) RGPDLARHGR and (B) REDVLIRHGK] are isolated and sequenced. These are analysed separately on the chip. Clone A is seen to bind specifically to the feature containing GCG - it is concluded that this clone is highly sequence-specific. Clone B lights up features with both GCG and GTG - this clone is bispecific. From the relative intensities of binding to the gradients of DNA on the chip it is concluded that the two clones have roughly equal affinity for the GCG site, and it is deduced that this affinity is in the nanomolar range.

Studying spacing requirements for zinc finger binding

DNA arrays are synthesised of the form 3' cctggctaactgaactATATAT-GCG-GGT-GCG-Nx-GCG-CAG-GCG-ATATAT 5', i.e. that contain variable nucleotide spacing (of 0 to 20 bp) between two 3-finger binding sites. Features are also included that contain one or the other binding site, but not both in a head to tail orientation as above. A 6-finger protein is constructed comprising a fusion between wild-type Zif268 three-fingers and a three-finger protein selected from the Choo and Klug library to bind GAC, linked by the linker H (zinc chelating)-LRQKDGERP-Y (hydrophobic core) where H and Y are the last and first structural elements of two adjacent fingers. The protein is applied to the chip and appreciable binding is seen to those features in which the spacing (Nx) is from 0 to 3 nucleotides, but no binding is observed to features where the spacing is greater than 7. It is concluded that the linker design restricts binding to short spacings between adjacent binding sites. From the relative intensities of binding to the gradients of DNA on the chip it is concluded that the protein binds to those features which contain both binding sites spaced by 0 to 3 bp much more tightly (100-fold tighter) than to features containing only one binding site - it is concluded that the protein shows high discrimination for the composite site relative to either half site.

Example 2: Construction of DNA-binding domains by phage display

A bipartite-complementary' system for the construction of DNA-binding domains by phage display may be used (Fig. 1). This system comprises two master libraries, Lib12 and Lib23, each of which encodes variants of a three-finger DNA-binding domain based on that of the transcription factor Zif268 (Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* 252, 809-817 (1991); Christy, B. A., Lau, L. F. & Nathans, D. A gene activated in mouse 3T3 cells by serum growth factors encodes a protein with "zinc finger" sequences. *Proc. Natl. Acad. Sci. USA* 85, 7857-7861 (1988).). The two libraries are complementary because Lib12 contains randomisations in all the base-contacting positions of F1 and certain base-contacting positions of F2, while Lib23 contains randomisations in the remaining base-contacting positions of F2 and all the base-contacting positions of F3 (Fig. 2a). The non-randomised DNA-contacting residues carry the nucleotide specificity of the parental Zif268 DNA-binding domain.

The design of the bipartite system features at least two modifications to the conventional zinc finger engineering strategies. As described above, each library contains members that are randomised in the α-helical DNA-contacting residues from more than one zinc finger. We have shown that the simultaneous randomisation of positions from adjacent fingers results in selected zinc finger pairs that can achieve comprehensive DNA recognition, i.e. bind DNA without significant sequence limitations.

The proteins produced by these libraries are therefore not limited to binding DNA sequences of the form GNNGNN..., as is the case with many prior art libraries (eg. 9, 13, 20).

The repertoire of randomisations does not encode all 20 amino acids, rather representing only those residues that most frequently function in sequence-specific DNA binding from the respective α-helical positions (Fig 2b). Excluding the residues that do not frequently

function in DNA recognition advantageously helps to reduce the library size and/or the 'noise' associated with non-specific binding members of the library.

Phage libraries for use in the present invention are prepared as follows.

Genes for the two zinc finger phage display libraries (Lib12 and Lib23) are assembled from synthetic DNA oligonucleotides by directional end-to-end ligation using short complementary DNA linkers. In order to include only the amino acids shown in Fig. 2b, a large number of appropriately randomised oligonucleotides (each encoding a subset of a few amino acids) are used in combinations to assemble the gene cassettes. These are amplified by PCR, digested with *Sfi*I and *Not*I endonucleases, and ligated into the phage vector Fd-Tet-SN (Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**, 809-817 (1991)).

E. coli TG1 cells are transformed with the recombinant vector by electroporation and plated onto TYE medium (1.5 % (w/v) agar, 1 % (w/v) Bactotryptone, 0.5 % (w/v) Bactoyeast extract, 0.8 % (w/v) NaCl) containing 15 µg/ml tetracycline.

The theoretical library sizes of Lib12 and Lib23 are approx. 4.9×10^6 and approx. 2.1×10^6, respectively (Fig. 2b).

Approximately twice these numbers of bacterial transformants are obtained for the respective libraries.

Example 3: Production of DNA-binding domains that target the HIV-1 promoter

Phage selections from the two master libraries described in Example 2 (Lib12 and Lib23) are performed using the generic DNA sequence 3'-HIJKLMGGCG-5' for Lib12, and 3'-GGCGMNOPQ-5' for Lib23, where the underlined bases are bound by the wild-type portion of the DNA-binding domain and each of the other letters represents any given nucleotide (Fig. 2a).

A number of sites in the well-characterised promoter of HIV-1 are targeted.

In this example, the two zinc finger libraries (Lib12 and Lib23) are subjected to selection in parallel, the nucleotide sequences used (ie. HIJKL/MNOPQ) being from HIV-1 between positions –80 and +60 (see Table 1/Fig. 3).

Tetracycline resistant bacterial colonies are transferred to 2 x TY liquid medium (16 g/litre Bactotryptone, 10 g/litre Bactoyeast extract, 5 g/litre NaCl) containing 50 μM $ZnCl_2$ and 15 μg/ml tetracycline, and cultured overnight at 30°C in a shaking incubator.

Cleared culture supernatant containing phage particles is obtained by centrifuging at 300 g for 5 minutes.

One picomole of biotinylated DNA target site is bound to streptavidin-coated tubes (Roche), in 50 μl PBS containing 50 μM $ZnCl_2$. Bacterial culture supernatant containing phage is diluted 1:10 in selection buffer (PBS containing 50 μM $ZnCl_2$, 2 % (w/v) fat-free dried milk (Marvel), 1 % (v/v) Tween, 20 mg/ml sonicated salmon sperm DNA), and 1 ml is applied to each tube. Binding reactions are incubated for 1 hour at 20°C, after which the tubes are emptied and washed 20 times with PBS containing 50 μM $ZnCl_2$, 2 % (w/v) fat-free dried milk (Marvel) and 1 % (v/v) Tween.

Retained phage are eluted in 0.1 M triethylamine and neutralised with an equal volume of 1 M Tris-HCl (pH 7.4). Logarithmic-phase *E. coli* TG1 are infected with eluted phage, and cultured overnight at 30°C in 2 ∞ TY medium containing 50 μM $ZnCl_2$ and 15 μg/ml tetracycline, to amplify phage for further rounds of selection.

After 5 rounds of selection, *E. coli* TG1 infected with selected phage are plated and individual colonies are picked and cultured in liquid medium to prepare phage for ELISA DNA-binding assays (Choo, Y. & Klug, A. Selection of DNA binding sites for zinc fingers

using rationally randomised DNA reveals coded interactions. *Proc. Natl. Acad. Sci. U.S.A.* 91, 11168-11172 (1994); Example 4).

Clones which recognise their target site may be retained for subsequent recombination of the two complementary halves recovered from Lib12 and Lib23 to produce molecules having high affinity for the HIV-1 promoter.

Eight DNA-binding domains are produced (Table 1, clones A-G; Clone H (HIV A') binds 5'-GCC TGG G(T/C)G-3' having the sequences F1-RSDVLTR; F2-RSDHLTT; F3-DYSVRKR).

Six (clones B-G) are engineered according to the full 'bipartite' protocol, while one protein (clone A) is derived directly by selection from Lib23. This illustrates a further use of the master libraries, namely to select zinc finger domains that bind DNA sequences containing the motif 5'-GCGG-3' or 5'-GGCG-3'.

Four proteins have binding sites that are dispersed upstream of the transcription initiation site (clones A-D), including two that flank the TATA box (clones C-D). Another three proteins bind to a cluster of sites at the beginning of the ORF, within the coding region for TAR (clones E-G). Clone H (HIV A') binds between the sites for HIV A and HIV B.

As the randomisations in the master libraries are restricted to amino acids with validated roles in DNA recognition, many of the recombinant DNA-binding domains make use of contacts that are consistent with the zinc finger-DNA 'recognition code' (Choo, Y. & Klug, A. Physical basis of a protein-DNA recognition code. *Curr. Opin. Str. Biol.* 7, 117-125 (1997).): e.g. the well-known RXD motif found at the N-terminus of many zinc finger α-helices is selected in clones A, B and G.

In summary, using our selection method we produced seven DNA-binding domains binding different loci in the genome of HIV-1 between positions –80 and +60 (Table 1).

Example 4: ELISA DNA Binding Assays

As noted above, the immobilised DNA library of the present invention may be used to verify the binding ability of rationally designed zinc fingers, or they may be used to screen for zinc fingers having specificity for one or more DNA sequences, or to determine the preferred base recognition specificity of a zinc finger. The binding specificity of the zinc finger sequences to a particular sequence or sequences within the immobilised library may be determined by any suitable binding assay as known in the art, for example, an ELISA assay as follows:

Equipment and reagents

- Sterile, round-bottom, 200 µl, 96-well plates for tissue culture (Costar, Corning USA)
- 2 x TY (Bacto tryptone, 16.0 g/l; Bacto yeast extract, 10.0 g/l; NaCl, 5.0 g/l)
- Tetracycline
- Zinc chloride, 1 M
- Streptavidin-coated microtitre well plates (Roche).
- PBS (10 x stock solution: NaCl, 80 g/l;KCl, 2g/l; $Na_2HPO_4.7H_2O$ 11.5 g/l;KH_2PO_4, 2 g/l)
- Fat-free freeze-dried milk (Marvel; Premier Brands UK Ltd.)
- Tween-20
- Sonicated salmon sperm DNA (10mg/ml)
- Horseradish peroxidase-conjugated anti-M13 IgG (Pharmacia Biotech)
- ELISA developer solution [0.1 M Na (CH_3.COO), pH 5.5; 3', 3', 5' 5'-tetramethylbenzidine (TMB; Sigma), 0.5 mg/ml; dimethyl sulphoxide (DMSO), 1% (v/v); H_2O_2, 0.05% (v/v)]
- Sulphuric acid, 1 M
- ELISA plate reader

Method

1. Pick single bacterial colonies containing phage clones derived from library selections. Use a sterile toothpick to transfer colonies to wells in sterile round-bottom plates containing 150 µl of 2 x TY µg/ml tetracycline. As a positive control, use one well to grow phage displaying the wild-type DNA-binding domain.

   Certain nucleic acid-binding domains may require supplements to the growth medium. Zinc fingers, for example, are stabilised by 50 µM$ZnCl_2$ in all media and ELISA binding and wash buffers. Incubate plates with orbital mixing at 250 rpm, for 16 hours at 30°C.

2. Add biotinylated nucleic acid target sites (typically between 0 – 5 pmol) in 50 µl PBS to streptavidin-coated microtitre wells (Roche). For the positive control, add an appropriate amount of the wild-type binding site to one well. Use a negative control well, containing PBS only, to measure the ELISA background. Bind DNA sites for 15 minutes at 20°C.
3. To each well, add 150 µl of PBS containing 4% (w/v) Marvel as a blocking agent. Leave blocking reaction for 1 hour at 20°C.
4. Prepare phage supernatant by centrifuging the 96-well culture plates at 3700 g for 15 minutes, in an appropriate swinging-bucket centrifuge.
5. Dilute phage supernatant 1:10 in 1 ml PBS containing 2% (w/v) Marvel, 1% (v/v) Tween-20 and 20 µg/ml sonicated salmon sperm DNA.
6. Discard blocking solution from the nucleic acid-coated wells and apply 50 µl of the diluted phage supernatant solution. Incubate for 1 hour at 20°C.
7. Discard the binding mixture and wash the well 7 times with 200 µl PBS, containing 1% (v/v) Tween-20. Wash a further 3 times with 200 µl PBS alone.
8. To each well, add 50 µl of PBS containing 2% (v/v) Marvel and a 1:5000 dilution of horseradish peroxidase (HRP)- conjugated anti-M13 IgG antibody (Pharmacia Biotech). Incubate at 20°C for 1 hour.

9. Discard the antibody binding mixture and wash the wells 3 times with 200 µl PBS, containing 0.05% (v/v) Tween-20. Wash a further 3 times with 200 µl PBS alone.

10. Develop the ELISA using 100 µl of HRP substrate such as the TMB-based ELISA developer solution described above. Stop the colorimetric reaction after approximately 5 minutes; for TMB add 100 µl of 1 H H_2SO_4. Quantitate the ELISA signals immediately using a spectrophotometer fitted a microtitre plate reader.

Although the protocol recited above relates to phage clones expressing zinc fingers which have been selected, the protocol may readily be adapted to assay interactions between specific zinc finger polypeptides and DNA substrates.

The ELISA DNA binding assay described above may be used to determine the binding specificity of a particular zinc finger, or a series of zinc fingers. Similarly, either a single DNA sequence or a series of DNA sequences may be tested.

Figure 1 shows the results of such an ELISA assay. Seven zinc finger DNA-binding domains are designed to bind sequences in the HIV-1 promoter. The seven constructs and their respective binding sites are labelled A-G, and each clone is tested using all seven DNA sequences. Binding of zinc fingers to 0.4 pmol DNA per 50µl well is plotted vertically from phage ELISA absorbance readings (A_{450}-A_{650}). As can be seen from the figure, strong and specific binding of each zinc finger is only observed to the DNA sequence against which it has been designed. See also Table 1.

All publications mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described methods and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention which are obvious to those skilled in molecular biology or related fields are intended to be within the scope of the following claims.

CLAIMS

1. A library of DNA sequences consisting of 4^N sequences, where N is greater than or equal to three, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of a DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate.

2. A method for designing a zinc finger polypeptide having specificity for a particular DNA sequence comprising a contiguous sequence of N nucleotides, where N is greater than or equal to three, which method comprises:

(i) providing a zinc finger polypeptide, preferably by designing using a rational design method or by selection from a library;

(ii) producing the polypeptide;

(iii) determining the sequence specificity for the polypeptide by contacting a library of DNA sequences with the polypeptide and identifying the sequence or sequences with which the polypeptide binds to with greatest affinity;

(iv) if the sequence or sequences identified in step (iii) are not the desired sequences, making modifications to the amino acid sequence of the polypeptide, preferably based on rational design or by selection from a library, and repeating steps (ii) and (iii),

wherein the library of DNA sequences consists of 4^N sequences, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of the DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate.

3. A method for isolating a zinc finger polypeptide having specificity for a particular DNA sequence comprising a contiguous sequence of N nucleotides, where N is greater than or equal to three, which method comprises:

(i) contacting a library of carrier organisms which express on their surface a zinc finger polypeptide comprising variations in the amino acid sequence of the zinc finger DNA binding domain, with a library of DNA sequences; and

(ii) selecting those carrier organisms which express a zinc finger polypeptide that binds to the particular DNA sequence; and

(iii) optionally repeating selection steps (i) and (ii) with those carrier organisms selected in step (ii),

wherein the library of DNA sequences consist of 4^N sequences, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of the DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate.

4. A method for determining the preferred base recognition specificity of a zinc finger polypeptide, which method comprises contacting a library of DNA sequences with the polypeptide, measuring the affinity with which the polypeptide binds to each of the sequences, and optionally ranking the sequences in order of the affinity with which the polypeptide binds,

wherein the library of DNA sequences consist of 4^N sequences, each sequence varying from the other sequences by comprising a different one of the 4^N possible permutations of the DNA sequence of length N, wherein the library of DNA sequences is immobilised on a solid substrate.

5. A library according to Claim 1, or a method according to any of Claims 2 to 4, in which each sequence of the library occupies a discrete position on the solid substrate.

6. Use of a library according to Claim 1 in a method for designing a zinc finger polypeptide having specificity for a particular DNA sequence.

7. Use of a library according to Claim 1 in a method for isolating a zinc finger polypeptide having specificity for a particular DNA sequence.

8. Use of a library according to Claim 1 in a method for determining the preferred base recognition specificity of a zinc finger polypeptide.

9. A library according to Claim 1, a method according to any of Claims 2 to 5, or a use according to any of Claims 6 to 8, in which the library is divided into two or more sub-libraries, in which each sub-library occupies a discrete position on the solid substrate.

10. A library, method or a use according to Claim 9, in which for any one sub-library one base in the DNA sequence of length N is defined and the other N-1 bases are randomised.

11. A sub-library according to Claim 9 or 10.

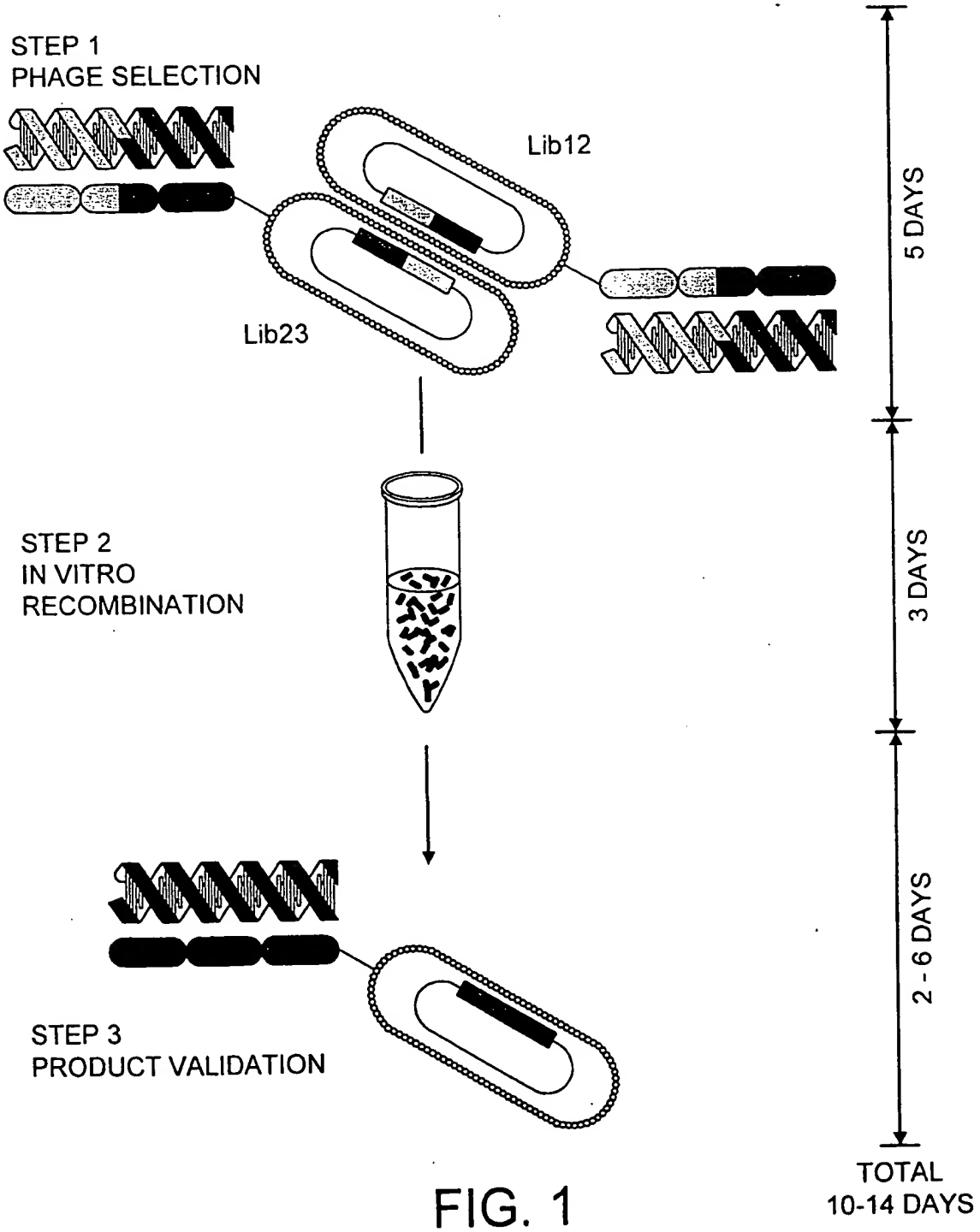
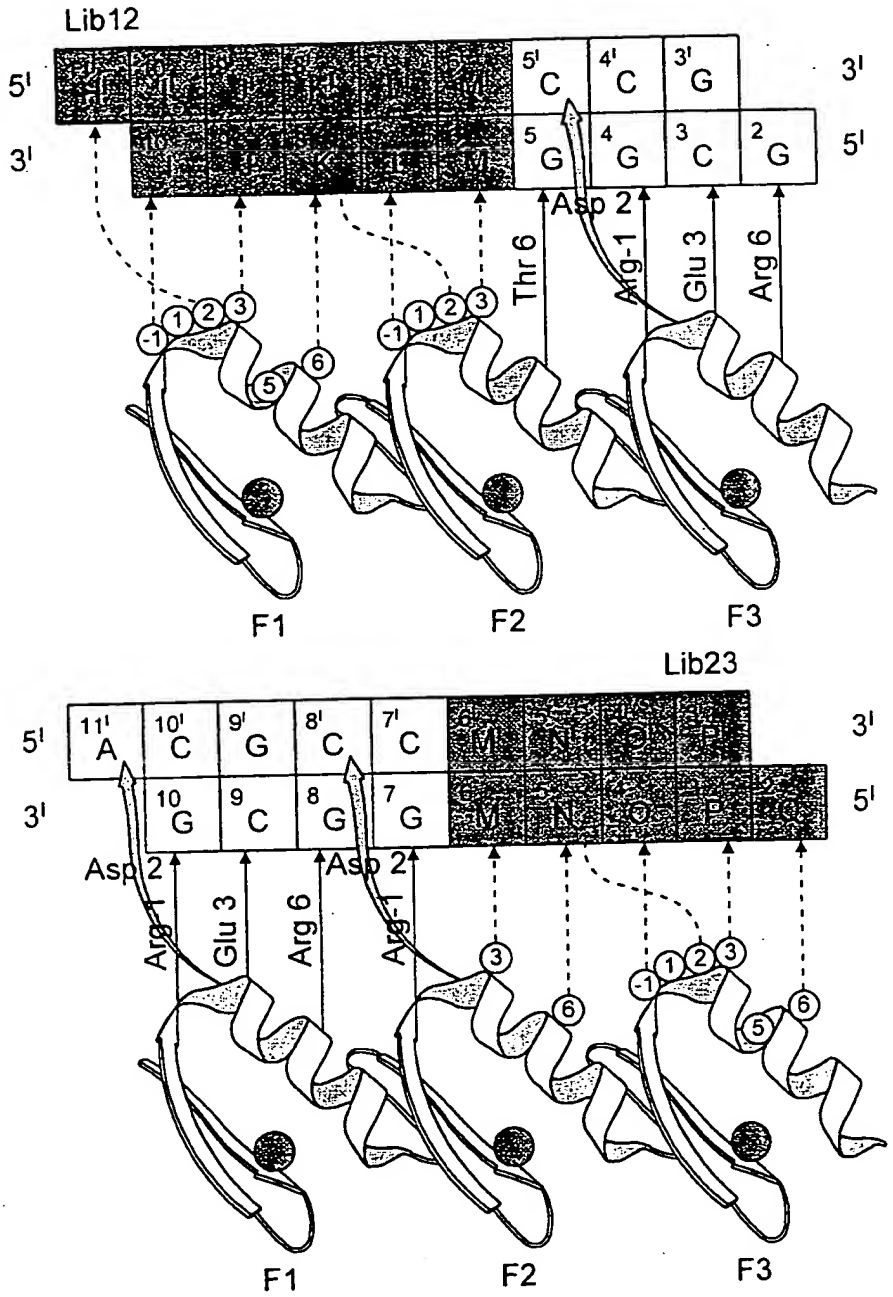

STEP 1
PHAGE SELECTION
Lib12
Lib23
STEP 2
IN VITRO
RECOMBINATION
STEP 3
PRODUCT VALIDATION
5 DAYS
3 DAYS
2 - 6 DAYS
TOTAL
10-14 DAYS

FIG. 1

a



b

-1123456	-1123456	-1123456
ASAALIA	ASAALSA	AKAARIA
D DD TE	DRDD E	DNDD KE
H HH K	H HH K	HRHH TK
N KN N	N KN N	NSKN N
Q NS Q	Q NS Q	Q NS Q
R QT R	R QT R	R QT R
T RV T	T RV T	T RV T
S V	S V	S V

FIG. 2

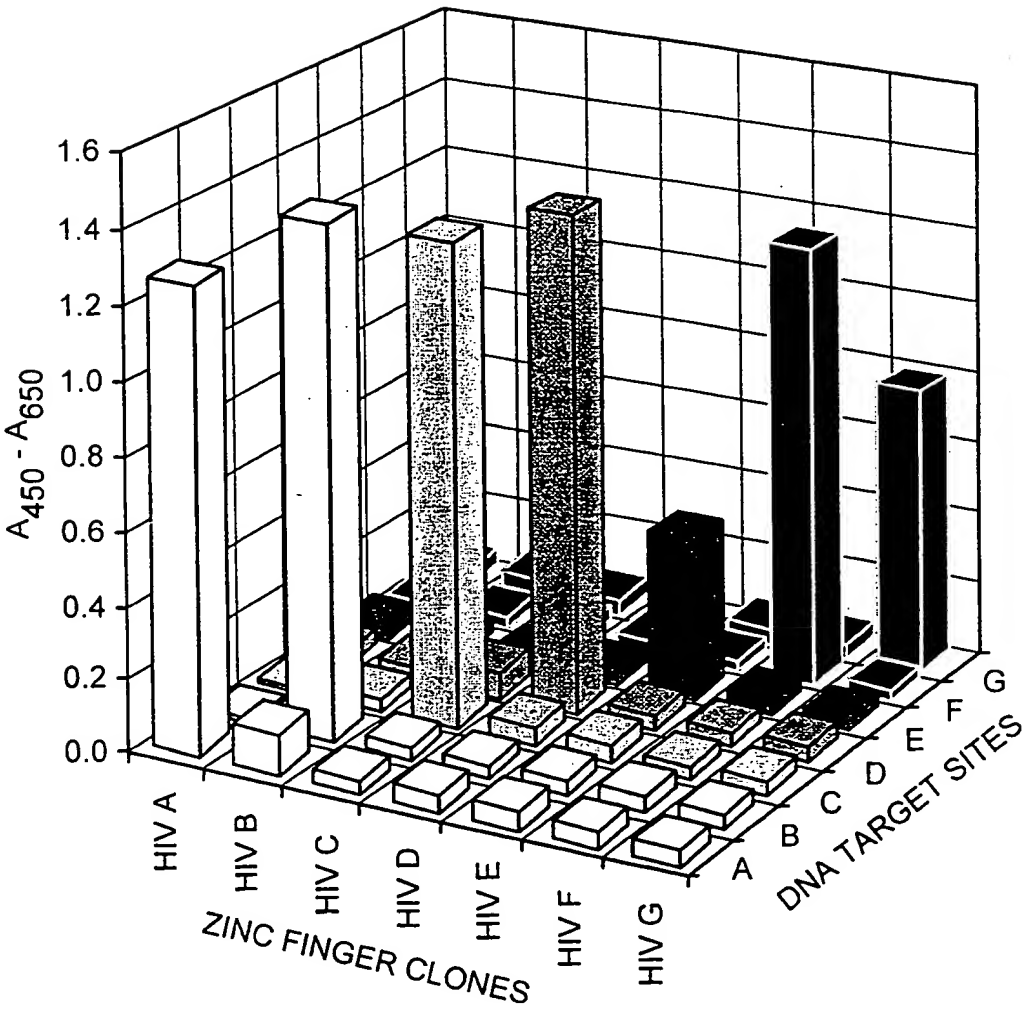

1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$A_{450} - A_{650}$
HIV A
HIV B
HIV C
HIV D
HIV E
HIV F
HIV G
ZINC FINGER CLONES
A
B
C
D
E
F
G
DNA TARGET SITES

FIG. 3

Clone	DNA target sequence[a]	Position in LTR	Zinc finger sequence[b]			K_d / nM[c]
	F1 F2 F3		F1	F2	F3	
	3'-**H IJK LMN** OPQ-5'		-1123456	-1123456	-1123456	
A	T GCG GAG GGA	-79	RSDELTR	RSDNLST	RRDHRTT	1.2±0.2
B	**G AGG GGT** CAG	-58	**DSAHLTR**	**RSDH**LST	DSANRTK	1.0±0.1
C	**T ACG TCG** TAG	-36	**ASADLTR**	**NRSD**LSR	TSSNRKK	13.7±3.6
D	**T TCG TCG** ACG	-22	**HSSDLTR**	**QSSD**LSK	QNATRKR	4.0±0.6
E	**T CCG AGT** CTA	+22	**DSSSLTK**	**QSAH**LST	DSSSRTK	36.6±15.0
F	**T CTC TCG** AGG	+33	**ASDDLTQ**	**RSSD**LSR	QSAHRTK	13.3±4.8
G	**G GAT CAA** TCG	+44	**RSDALIQ**	**DRAN**LST	ASSTRTK	40.3±14.6

TABLE 1